APPLICATION FOR UNITED STATES PATENT

INVENTORS: ALAN J. LIPTON
THOMAS M. STRAT
PÉTER L. VENETIANER
MARK C. ALLMEN
WILLIAM E. SEVERSON
NIELS HAERING
ANDREW J. CHOSAK
ZHONG ZHANG
MATTHEW F. FRAZIER
JAMES S. SFEKAS
TASUKI HIRATA
JOHN CLARK

TITLE: VIDEO SURVEILLANCE SYSTEM EMPLOYING VIDEO PRIMITIVES

ATTORNEYS' ADDRESS:

VENABLE
1201 New York Avenue, N.W., Suite 1000
Washington, D.C. 20005-3917
Telephone: (202) 962-4800
Telefax: (202) 962-8300

ADDRESS FOR U.S.P.T.O. CORRESPONDENCE:

VENABLE
Post Office Box 34385
Washington, D.C. 20043-9998

ATTORNEY DOCKET NO.:

37112-175340

VIDEO SURVEILLANCE SYSTEM EMPLOYING VIDEO PRIMITIVES

CROSS-REFERENCE TO RELATED APPLICATIONS

[1] This application claims the priority of U.S. Patent Application No. 09/694,712 filed October 24, 2000, which is incorporated herein by reference.

BACKGROUND OF THE INVENTION

Field of the Invention

[2] The invention relates to a system for automatic video surveillance employing video primitives.

References

[3] For the convenience of the reader, the references referred to herein are listed below. In the specification, the numerals within brackets refer to respective references. The listed references are incorporated herein by reference.

[4] The following references describe moving target detection:

[5] {1} A. Lipton, H. Fujiyoshi and R. S. Patil, "Moving Target Detection and Classification from Real-Time Video," Proceedings of IEEE WACV '98, Princeton, NJ, 1998, pp. 8-14.

[6] {2} W.E.L. Grimson, et al., "Using Adaptive Tracking to Classify and Monitor Activities in a Site", CVPR, pp. 22-29, June 1998.

[7] {3} A.J. Lipton, H. Fujiyoshi, R.S. Patil, "Moving Target Classification and Tracking from Real-time Video," IUW, pp. 129-136, 1998.

[8] {4} T.J. Olson and F.Z. Brill, "Moving Object Detection and Event Recognition Algorithm for Smart Cameras," IUW, pp. 159-175, May 1997.

[9] The following references describe detecting and tracking humans:

[10] {5} A. J. Lipton, "Local Application of Optical Flow to Analyse Rigid Versus Non-Rigid Motion," International Conference on Computer Vision, Corfu, Greece, September 1999.

[11] {6} F. Bartolini, V. Cappellini, and A. Mecocci, "Counting people getting in and out of a bus by real-time image-sequence processing," IVC, 12(1):36-41, January 1994.

[12] {7} M. Rossi and A. Bozzoli, "Tracking and counting moving people," ICIP94, pp. 212-216, 1994.

[13] {8} C.R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," Vismod, 1995.

[14] {9} L. Khoudour, L. Duvieubourg, J.P. Deparis, "Real-Time Pedestrian Counting by Active Linear Cameras," JEI, 5(4):452-459, October 1996.

[15] {10} S. Ioffe, D.A. Forsyth, "Probabilistic Methods for Finding People," IJCV, 43(1):45-68, June 2001.

[16] {11} M. Isard and J. MacCormick, "BraMBLe: A Bayesian Multiple-Blob Tracker," ICCV, 2001.

[17] The following references describe blob analysis:

[18] {12} D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," CVIU, 73(1):82-98, January 1999.

[19] {13} Niels Haering and Niels da Vitoria Lobo, "Visual Event Detection," Video Computing Series, Editor Mubarak Shah, 2001.

[20] The following references describe blob analysis for trucks, cars, and people:

[21] {14} Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver, Enomoto, and Hasegawa, "A System for Video Surveillance and Monitoring: VSAM Final Report," Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, May 2000.

[22] {15} Lipton, Fujiyoshi, and Patil, "Moving Target Classification and Tracking from Real-time Video," 98 Darpa IUW, Nov. 20-23, 1998.

[23] The following reference describes analyzing a single-person blob and its contours:

[24] {16} C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland. "Pfinder: Real-Time Tracking of the Human Body," PAMI, vol 19, pp. 780-784, 1997.

[25] The following reference describes internal motion of blobs, including any motion-based segmentation:

[26] {17} M. Allmen and C. Dyer, "Long--Range Spatiotemporal Motion Understanding Using Spatiotemporal Flow Curves," Proc. IEEE CVPR, Lahaina, Maui, Hawaii, pp. 303-309, 1991.

[27] {18} L. Wixson, "Detecting Salient Motion by Accumulating Directionally Consistent Flow", IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pp. 774-781, Aug, 2000.

Background of the Invention

[28] Video surveillance of public spaces has become extremely widespread and accepted by the general public. Unfortunately, conventional video surveillance systems produce such prodigious volumes of data that an intractable problem results in the analysis of video surveillance data.

[29] A need exists to reduce the amount of video surveillance data so analysis of the video surveillance data can be conducted.

[30] A need exists to filter video surveillance data to identify desired portions of the video surveillance data.

SUMMARY OF THE INVENTION

[31] An object of the invention is to reduce the amount of video surveillance data so analysis of the video surveillance data can be conducted.

[32] An object of the invention is to filter video surveillance data to identify desired portions of the video surveillance data.

[33] An object of the invention is to produce a real time alarm based on an automatic detection of an event from video surveillance data.

[34] An object of the invention is to integrate data from surveillance sensors other than video for improved searching capabilities.

[35] An object of the invention is to integrate data from surveillance sensors other than video for improved event detection capabilities

[36] The invention includes an article of manufacture, a method, a system, and an apparatus for video surveillance.

[37] The article of manufacture of the invention includes a computer-readable medium comprising software for a video surveillance system, comprising code segments for operating the video surveillance system based on video primitives.

[38] The article of manufacture of the invention includes a computer-readable medium comprising software for a video surveillance system, comprising code segments for accessing

archived video primitives, and code segments for extracting event occurrences from accessed archived video primitives.

[39] The system of the invention includes a computer system including a computer-readable medium having software to operate a computer in accordance with the invention.

[40] The apparatus of the invention includes a computer including a computer-readable medium having software to operate the computer in accordance with the invention.

[41] The article of manufacture of the invention includes a computer-readable medium having software to operate a computer in accordance with the invention.

[42] Moreover, the above objects and advantages of the invention are illustrative, and not exhaustive, of those that can be achieved by the invention. Thus, these and other objects and advantages of the invention will be apparent from the description herein, both as embodied herein and as modified in view of any variations which will be apparent to those skilled in the art.

Definitions

[43] A "video" refers to motion pictures represented in analog and/or digital form. Examples of video include: television, movies, image sequences from a video camera or other observer, and computer-generated image sequences.

[44] A "frame" refers to a particular image or other discrete unit within a video.

[45] An "object" refers to an item of interest in a video. Examples of an object include: a person, a vehicle, an animal, and a physical subject.

[46] An "activity" refers to one or more actions and/or one or more composites of actions of one or more objects. Examples of an activity include: entering; exiting; stopping; moving; raising; lowering; growing; and shrinking.

[47] A "location" refers to a space where an activity may occur. A location can be, for example, scene-based or image-based. Examples of a scene-based location include: a public space; a store; a retail space; an office; a warehouse; a hotel room; a hotel lobby; a lobby of a building; a casino; a bus station; a train station; an airport; a port; a bus; a train; an airplane; and a ship. Examples of an image-based location include: a video image; a line in a video image; an area in a video image; a rectangular section of a video image; and a polygonal section of a video image.

[48] An "event" refers to one or more objects engaged in an activity. The event may be referenced with respect to a location and/or a time.

[49] A "computer" refers to any apparatus that is capable of accepting a structured input, processing the structured input according to prescribed rules, and producing results of the processing as output. Examples of a computer include: a computer; a general purpose computer; a supercomputer; a mainframe; a super mini-computer; a mini-computer; a workstation; a micro-computer; a server; an interactive television; a hybrid combination of a computer and an interactive television; and application-specific hardware to emulate a computer and/or software. A computer can have a single processor or multiple processors, which can operate in parallel and/or not in parallel. A computer also refers to two or more computers connected together via a network for transmitting or receiving information between the computers. An example of such a computer includes a distributed computer system for processing information via computers linked by a network.

[50] A "computer-readable medium" refers to any storage device used for storing data accessible by a computer. Examples of a computer-readable medium include: a magnetic hard disk; a floppy disk; an optical disk, such as a CD-ROM and a DVD; a magnetic tape; a memory chip; and a carrier wave used to carry computer-readable electronic data, such as those used in transmitting and receiving e-mail or in accessing a network.

[51] "Software" refers to prescribed rules to operate a computer. Examples of software include: software; code segments; instructions; computer programs; and programmed logic.

[52] A "computer system" refers to a system having a computer, where the computer comprises a computer-readable medium embodying software to operate the computer.

[53] A "network" refers to a number of computers and associated devices that are connected by communication facilities. A network involves permanent connections such as cables or temporary connections such as those made through telephone or other communication links. Examples of a network include: an internet, such as the Internet; an intranet; a local area network (LAN); a wide area network (WAN); and a combination of networks, such as an internet and an intranet.

BRIEF DESCRIPTION OF THE DRAWINGS

[54] Embodiments of the invention are explained in greater detail by way of the drawings, where the same reference numerals refer to the same features.

[55] Figure 1 illustrates a plan view of the video surveillance system of the invention.

[56] Figure 2 illustrates a flow diagram for the video surveillance system of the invention.

[57] Figure 3 illustrates a flow diagram for tasking the video surveillance system.

[58] Figure 4 illustrates a flow diagram for operating the video surveillance system.

[59] Figure 5 illustrates a flow diagram for extracting video primitives for the video surveillance system.

[60] Figure 6 illustrates a flow diagram for taking action with the video surveillance system.

[61] Figure 7 illustrates a flow diagram for semi-automatic calibration of the video surveillance system.

[62] Figure 8 illustrates a flow diagram for automatic calibration of the video surveillance system.

[63] Figure 9 illustrates an additional flow diagram for the video surveillance system of the invention.

[64] Figures 10-15 illustrate examples of the video surveillance system of the invention applied to monitoring a grocery store.

DETAILED DESCRIPTION OF THE INVENTION

[65] The automatic video surveillance system of the invention is for monitoring a location for, for example, market research or security purposes. The system can be a dedicated video surveillance installation with purpose-built surveillance components, or the system can be a retrofit to existing video surveillance equipment that piggybacks off the surveillance video feeds. The system is capable of analyzing video data from live sources or from recorded media. The system can have a prescribed response to the analysis, such as record data, activate an alarm mechanism, or active another sensor system. The system is also capable of integrating with

other surveillance system components. The system produces security or market research reports that can be tailored according to the needs of an operator and, as an option, can be presented through an interactive web-based interface, or other reporting mechanism.

[66] An operator is provided with maximum flexibility in configuring the system by using event discriminators. Event discriminators are identified with one or more objects (whose descriptions are based on video primitives), along with one or more optional spatial attributes, and/or one or more optional temporal attributes. For example, an operator can define an event discriminator (called a "loitering" event in this example) as a "person" object in the "automatic teller machine" space for "longer than 15 minutes" and "between 10:00 p.m. and 6:00 a.m."

[67] Although the video surveillance system of the invention draws on well-known computer vision techniques from the public domain, the inventive video surveillance system has several unique and novel features that are not currently available. For example, current video surveillance systems use large volumes of video imagery as the primary commodity of information interchange. The system of the invention uses video primitives as the primary commodity with representative video imagery being used as collateral evidence. The system of the invention can also be calibrated (manually, semi-automatically, or automatically) and thereafter automatically can infer video primitives from video imagery. The system can further analyze previously processed video without needing to reprocess completely the video. By analyzing previously processed video, the system can perform inference analysis based on previously recorded video primitives, which greatly improves the analysis speed of the computer system.

[68] As another example, the system of the invention provides unique system tasking. Using equipment control directives, current video systems allow a user to position video sensors

and, in some sophisticated conventional systems, to mask out regions of interest or disinterest. Equipment control directives are instructions to control the position, orientation, and focus of video cameras. Instead of equipment control directives, the system of the invention uses event discriminators based on video primitives as the primary tasking mechanism. With event discriminators and video primitives, an operator is provided with a much more intuitive approach over conventional systems for extracting useful information from the system. Rather than tasking a system with an equipment control directives, such as "camera A pan 45 degrees to the left," the system of the invention can be tasked in a human-intuitive manner with one or more event discriminators based on video primitives, such as "a person enters restricted area A."

[69] Using the invention for market research, the following are examples of the type of video surveillance that can be performed with the invention: counting people in a store; counting people in a part of a store; counting people who stop in a particular place in a store; measuring how long people spend in a store; measuring how long people spend in a part of a store; and measuring the length of a line in a store.

[70] Using the invention for security, the following are examples of the type of video surveillance that can be performed with the invention: determining when anyone enters a restricted area and storing associated imagery; determining when a person enters an area at unusual times; determining when changes to shelf space and storage space occur that might be unauthorized; determining when passengers aboard an aircraft approach the cockpit; determining when people tailgate through a secure portal; determining if there is an unattended bag in an airport; and determining if there is a theft of an asset.

[71] Figure 1 illustrates a plan view of the video surveillance system of the invention. A computer system 11 comprises a computer 12 having a computer-readable medium 13

embodying software to operate the computer 12 according to the invention. The computer system 11 is coupled to one or more video sensors 14, one or more video recorders 15, and one or more input/output (I/O) devices 16. The video sensors 14 can also be optionally coupled to the video recorders 15 for direct recording of video surveillance data. The computer system is optionally coupled to other sensors 17.

[72] The video sensors 14 provide source video to the computer system 11. Each video sensor 14 can be coupled to the computer system 11 using, for example, a direct connection (e.g., a firewire digital camera interface) or a network. The video sensors 14 can exist prior to installation of the invention or can be installed as part of the invention. Examples of a video sensor 14 include: a video camera; a digital video camera; a color camera; a monochrome camera; a camera; a camcorder, a PC camera; a webcam; an infra-red video camera; and a CCTV camera.

[73] The video recorders 15 receive video surveillance data from the computer system 11 for recording and/or provide source video to the computer system 11. Each video recorder 15 can be coupled to the computer system 11 using, for example, a direct connection or a network. The video recorders 15 can exist prior to installation of the invention or can be installed as part of the invention. Examples of a video recorder 15 include: a video tape recorder; a digital video recorder; a video disk; a DVD; and a computer-readable medium.

[74] The I/O devices 16 provide input to and receive output from the computer system 11. The I/O devices 16 can be used to task the computer system 11 and produce reports from the computer system 11. Examples of I/O devices 16 include: a keyboard; a mouse; a stylus; a monitor; a printer; another computer system; a network; and an alarm.

[75] The other sensors 17 provide additional input to the computer system 11. Each other sensor 17 can be coupled to the computer system 11 using, for example, a direct connection or a network. The other sensors 17 can exit prior to installation of the invention or can be installed as part of the invention. Examples of another sensor 17 include: a motion sensor; an optical tripwire; a biometric sensor; and a card-based or keypad-based authorization system. The outputs of the other sensors 17 can be recorded by the computer system 11, recording devices, and/or recording systems.

[76] Figure 2 illustrates a flow diagram for the video surveillance system of the invention. Various aspects of the invention are exemplified with reference to Figures 10-15, which illustrate examples of the video surveillance system of the invention applied to monitoring a grocery store.

[77] In block 21, the video surveillance system is set up as discussed for Figure 1. Each video sensor 14 is orientated to a location for video surveillance. The computer system 11 is connected to the video feeds from the video equipment 14 and 15. The video surveillance system can be implemented using existing equipment or newly installed equipment for the location.

[78] In block 22, the video surveillance system is calibrated. Once the video surveillance system is in place from block 21, calibration occurs. The result of block 22 is the ability of the video surveillance system to determine an approximate absolute size and speed of a particular object (e.g., a person) at various places in the video image provided by the video sensor. The system can be calibrated using manual calibration, semi-automatic calibration, and automatic calibration. Calibration is further described after the discussion of block 24.

[79] In block 23 of Figure 2, the video surveillance system is tasked. Tasking occurs after calibration in block 22 and is optional. Tasking the video surveillance system involves specifying one or more event discriminators. Without tasking, the video surveillance system operates by detecting and archiving video primitives and associated video imagery without taking any action, as in block 45 in Figure 4.

[80] Figure 3 illustrates a flow diagram for tasking the video surveillance system to determine event discriminators. An event discriminator refers to one or more objects optionally interacting with one or more spatial attributes and/or one or more temporal attributes. An event discriminator is described in terms of video primitives. A video primitive refers to an observable attribute of an object viewed in a video feed. Examples of video primitives include the following: a classification; a size; a shape; a color; a texture; a position; a velocity; a speed; an internal motion; a motion; a salient motion; a feature of a salient motion; a scene change; a feature of a scene change; and a pre-defined model.

[81] A classification refers to an identification of an object as belonging to a particular category or class. Examples of a classification include: a person; a dog; a vehicle; a police car; an individual person; and a specific type of object.

[82] A size refers to a dimensional attribute of an object. Examples of a size include: large; medium; small; flat; taller than 6 feet; shorter than 1 foot; wider than 3 feet; thinner than 4 feet; about human size; bigger than a human; smaller than a human; about the size of a car; a rectangle in an image with approximate dimensions in pixels; and a number of image pixels.

[83] A color refers to a chromatic attribute of an object. Examples of a color include: white; black; grey; red; a range of HSV values; a range of YUV values; a range of RGB values; an average RGB value; an average YUV value; and a histogram of RGB values.

[84] A texture refers to a pattern attribute of an object. Examples of texture features include: self-similarity; spectral power; linearity; and coarseness.

[85] An internal motion refers to a measure of the rigidity of an object. An example of a fairly rigid object is a car, which does not exhibit a great amount of internal motion. An example of a fairly non-rigid object is a person having swinging arms and legs, which exhibits a great amount of internal motion.

[86] A motion refers to any motion that can be automatically detected. Examples of a motion include: appearance of an object; disappearance of an object; a vertical movement of an object; a horizontal movement of an object; and a periodic movement of an object.

[87] A salient motion refers to any motion that can be automatically detected and can be tracked for some period of time. Such a moving object exhibits apparently purposeful motion. Examples of a salient motion include: moving from one place to another; and moving to interact with another object.

[88] A feature of a salient motion refers to a property of a salient motion. Examples of a feature of a salient motion include: a trajectory; a length of a trajectory in image space; an approximate length of a trajectory in a three-dimensional representation of the environment; a position of an object in image space as a function of time; an approximate position of an object in a three-dimensional representation of the environment as a function of time; a duration of a trajectory; a velocity (e.g., speed and direction) in image space; an approximate velocity (e.g., speed and direction) in a three-dimensional representation of the environment; a duration of time at a velocity; a change of velocity in image space; an approximate change of velocity in a three-dimensional representation of the environment; a duration of a change of velocity; cessation of motion; and a duration of cessation of motion. A velocity refers to the speed and direction of an

object at a particular time. A trajectory refers a set of (position, velocity) pairs for an object for as long as the object can be tracked or for a time period.

[89] A scene change refers to any region of a scene that can be detected as changing over a period of time. Examples of a scene change include: an stationary object leaving a scene; an object entering a scene and becoming stationary; an object changing position in a scene; and an object changing appearance (e.g. color, shape, or size).

[90] A feature of a scene change refers to a property of a scene change. Examples of a feature of a scene change include: a size of a scene change in image space; an approximate size of a scene change in a three-dimensional representation of the environment; a time at which a scene change occurred; a location of a scene change in image space; and an approximate location of a scene change in a three-dimensional representation of the environment.

[91] A pre-defined model refers to an *a priori* known model of an object. Examples of a pre-defined include: an adult; a child; a vehicle; and a semi-trailer.

[92] In block 31, one or more objects types of interests are identified in terms of video primitives or abstractions thereof. Examples of one or more objects include: an object; a person; a red object; two objects; two persons; and a vehicle.

[93] In block 32, one or more spatial areas of interest are identified. An area refers to one or more portions of an image from a source video or a spatial portion of a scene being viewed by a video sensor. An area also includes a combination of areas from various scenes and/or images. An area can be an image-based space (e.g., a line, a rectangle, a polygon, or a circle in a video image) or a three-dimensional space (e.g., a cube, or an area of floor space in a building).

[94] Figure 12 illustrates identifying areas along an aisle in a grocery store. Four areas are identified: coffee; soda promotion; chips snacks; and bottled water. The areas are identified via a point-and-click interface with the system.

[95] In block 33, one or more temporal attributes of interest are optionally identified. Examples of a temporal attribute include: every 15 minutes; between 9:00 p.m. to 6:30 a.m.; less than 5 minutes; longer than 30 seconds; over the weekend; and within 20 minutes of.

[96] In block 34, a response is optionally identified. Examples of a response includes the following: activating a visual and/or audio alert on a system display; activating a visual and/or audio alarm system at the location; activating a silent alarm; activating a rapid response mechanism; locking a door; contacting a security service; forwarding data (e.g., image data, video data, video primitives; and/or analyzed data) to another computer system via a network, such as the Internet; saving such data to a designated computer-readable medium; activating some other sensor or surveillance system; tasking the computer system 11 and/or another computer system; and directing the computer system 11 and/or another computer system.

[97] In block 35, one or more discriminators are identified by describing interactions between video primitives (or their abstractions), spatial areas of interest, and temporal attributes of interest. An interaction is determined for a combination of one or more objects identified in block 31, one or more spatial areas of interest identified in block 32, and one or more temporal attributes of interest identified in block 33. One or more responses identified in block 34 are optionally associated with each event discriminator.

[98] Examples of an event discriminator for a single object include: an object appears; a person appears; and a red object moves faster than 10m/s.

[99] Examples of an event discriminator for multiple objects include: two objects come together; a person exits a vehicle; and a red object moves next to a blue object.

[100] Examples of an event discriminator for an object and a spatial attribute include: an object crosses a line; an object enters an area; and a person crosses a line from the left.

[101] Examples of an event discriminator for an object and a temporal attribute include: an object appears at 10:00 p.m.; a person travels faster then 2m/s between 9:00 a.m. and 5:00 p.m.; and a vehicle appears on the weekend.

[102] Examples of an event discriminator for an object, a spatial attribute, and a temporal attribute include: a person crosses a line between midnight and 6:00 a.m.; and a vehicle stops in an area for longer than 10 minutes.

[103] An example of an event discriminator for an object, a spatial attribute, and a temporal attribute associated with a response include: a person enters an area between midnight and 6:00 a.m., and a security service is notified.

[104] In block 24 of Figure 2, the video surveillance system is operated. The video surveillance system of the invention operates automatically, detects and archives video primitives of objects in the scene, and detects event occurrences in real time using event discriminators. In addition, action is taken in real time, as appropriate, such as activating alarms, generating reports, and generating output. The reports and output can be displayed and/or stored locally to the system or elsewhere via a network, such as the Internet. Figure 4 illustrates a flow diagram for operating the video surveillance system.

[105] In block 41, the computer system 11 obtains source video from the video sensors 14 and/or the video recorders 15.

[106] In block 42, video primitives are extracted in real time from the source video. As an option, non-video primitives can be obtained and/or extracted from one or more other sensors 17 and used with the invention. The extraction of video primitives is illustrated with Figure 5.

[107] Figure 5 illustrates a flow diagram for extracting video primitives for the video surveillance system. Blocks 51 and 52 operate in parallel and can be performed in any order or concurrently. In block 51, objects are detected via movement. Any motion detection algorithm for detecting movement between frames at the pixel level can be used for this block. As an example, the three frame differencing technique can be used, which is discussed in {1}. The detected objects are forwarded to block 53.

[108] In block 52, objects are detected via change. Any change detection algorithm for detecting changes from a background model can be used for this block. An object is detected in this block if one or more pixels in a frame are deemed to be in the foreground of the frame because the pixels do not conform to a background model of the frame. As an example, a stochastic background modeling technique, such as dynamically adaptive background subtraction, can be used, which is described in {1} and U.S. Patent Application No. 09/694,712 filed October 24, 2000. The detected objects are forwarded to block 53.

[109] The motion detection technique of block 51 and the change detection technique of block 52 are complimentary techniques, where each technique advantageously addresses deficiencies in the other technique. As an option, additional and/or alternative detection schemes can be used for the techniques discussed for blocks 51 and 52. Examples of an additional and/or alternative detection scheme include the following: the Pfinder detection scheme for finding people as described in {8}; a skin tone detection scheme; a face detection scheme; and a model-

based detection scheme. The results of such additional and/or alternative detection schemes are provided to block 53.

[110] As an option, if the video sensor 14 has motion (e.g., a video camera that sweeps, zooms, and/or translates), an additional block can be inserted before blocks between blocks 51 and 52 to provide input to blocks 51 and 52 for video stabilization. Video stabilization can be achieved by affine or projective global motion compensation. For example, image alignment described in U.S. Patent Application No. 09/609,919, filed July 3, 2000, which is incorporated herein by reference, can be used to obtain video stabilization.

[111] In block 53, blobs are generated. In general, a blob is any object in a frame. Examples of a blob include: a moving object, such as a person or a vehicle; and a consumer product, such as a piece of furniture, a clothing item, or a retail shelf item. Blobs are generated using the detected objects from blocks 32 and 33. Any technique for generating blobs can be used for this block. An exemplary technique for generating blobs from motion detection and change detection uses a connected components scheme. For example, the morphology and connected components algorithm can be used, which is described in {1}.

[112] In block 54, blobs are tracked. Any technique for tracking blobs can be used for this block. For example, Kalman filtering or the CONDENSATION algorithm can be used. As another example, a template matching technique, such as described in {1}, can be used. As a further example, a multi-hypothesis Kalman tracker can be used, which is described in {5}. As yet another example, the frame-to-frame tracking technique described in U.S. Patent Application No. 09/694,712 filed October 24, 2000, can be used. For the example of a location being a grocery store, examples of objects that can be tracked include moving people, inventory items, and inventory moving appliances, such as shopping carts or trolleys.

[113] As an option, blocks 51-54 can be replaced with any detection and tracking scheme, as is known to those of ordinary skill. An example of such a detection and tracking scheme is described in {11}.

[114] In block 55, each trajectory of the tracked objects is analyzed to determine if the trajectory is salient. If the trajectory is insalient, the trajectory represents an object exhibiting unstable motion or represents an object of unstable size or color, and the corresponding object is rejected and is no longer analyzed by the system. If the trajectory is salient, the trajectory represents an object that is potentially of interest. A trajectory is determined to be salient or insalient by applying a salience measure to the trajectory. Techniques for determining a trajectory to be salient or insalient are described in {13} and {18}.

[115] In block 56, each object is classified. The general type of each object is determined as the classification of the object. Classification can be performed by a number of techniques, and examples of such techniques include using a neural network classifier {14} and using a linear discriminatant classifier {14}. Examples of classification are the same as those discussed for block 23.

[116] In block 57, video primitives are identified using the information from blocks 51-56 and additional processing as necessary. Examples of video primitives identified are the same as those discussed for block 23. As an example, for size, the system can use information obtained from calibration in block 22 as a video primitive. From calibration, the system has sufficient information to determine the approximate size of an object. As another example, the system can use velocity as measured from block 54 as a video primitive.

[117] In block 43, the video primitives from block 42 are archived. The video primitives can be archived in the computer-readable medium 13 or another computer-readable

medium. Along with the video primitives, associated frames or video imagery from the source video can be archived.

[118] In block 44, event occurrences are extracted from the video primitives using event discriminators. The video primitives are determined in block 42, and the event discriminators are determined from tasking the system in block 23. The event discriminators are used to filter the video primitives to determine if any event occurrences occurred. For example, an event discriminator can be looking for a "wrong way" event as defined by a person traveling the "wrong way" into an area between 9:00a.m. and 5:00p.m. The event discriminator checks all video primitives being generated according to Figure 5 and determines if any video primitives exist which have the following properties: a timestamp between 9:00a.m. and 5:00p.m., a classification of "person" or "group of people", a position inside the area, and a "wrong" direction of motion.

[119] In block 45, action is taken for each event occurrence extracted in block 44, as appropriate. Figure 6 illustrates a flow diagram for taking action with the video surveillance system.

[120] In block 61, responses are undertaken as dictated by the event discriminators that detected the event occurrences. The response, if any, are identified for each event discriminator in block 34.

[121] In block 62, an activity record is generated for each event occurrence that occurred. The activity record includes, for example: details of a trajectory of an object; a time of detection of an object; a position of detection of an object, and a description or definition of the event discriminator that was employed. The activity record can include information, such as video primitives, needed by the event discriminator. The activity record can also include

representative video or still imagery of the object(s) and/or area(s) involved in the event occurrence. The activity record is stored on a computer-readable medium.

[122] In block 63, output is generated. The output is based on the event occurrences extracted in block 44 and a direct feed of the source video from block 41. The output is stored on a computer-readable medium, displayed on the computer system 11 or another computer system, or forwarded to another computer system. As the system operates, information regarding event occurrences is collected, and the information can be viewed by the operator at any time, including real time. Examples of formats for receiving the information include: a display on a monitor of a computer system; a hard copy; a computer-readable medium; and an interactive web page.

[123] The output can include a display from the direct feed of the source video from block 41. For example, the source video can be displayed on a window of the monitor of a computer system or on a closed-circuit monitor. Further, the output can include source video marked up with graphics to highlight the objects and/or areas involved in the event occurrence.

[124] The output can include one or more reports for an operator based on the requirements of the operator and/or the event occurrences. Examples of a report include: the number of event occurrences which occurred; the positions in the scene in which the event occurrence occurred; the times at which the event occurrences occurred; representative imagery of each event occurrence; representative video of each event occurrence; raw statistical data; statistics of event occurrences (e.g., how many, how often, where, and when); and/or human-readable graphical displays.

[125] Figures 13 and 14 illustrate an exemplary report for the aisle in the grocery store of Figure 15. In Figures 13 and 14, several areas are identified in block 22 and are labeled

accordingly in the images. The areas in Figure 13 match those in Figure 12, and the areas in Figure 14 are different ones. The system is tasked to look for people who stop in the area.

[126] In Figure 13, the exemplary report is an image from a video marked-up to include labels, graphics, statistical information, and an analysis of the statistical information. For example, the area identified as coffee has statistical information of an average number of customers in the area of 2/hour and an average dwell time in the area as 5 seconds. The system determined this area to be a "cold" region, which means there is not much commercial activity through this region. As another example, the area identified as sodas has statistical information of an average number of customers in the area of 15/hour and an average dwell time in the area as 22 seconds. The system determined this area to be a "hot" region, which means there is a large amount of commercial activity in this region.

[127] In Figure 14, the exemplary report is an image from a video marked-up to include labels, graphics, statistical information, and an analysis of the statistical information. For example, the area at the back of the aisle has average number of customers of 14/hour and is determined to have low traffic. As another example, the area at the front of the aisle has average number of customers of 83/hour and is determined to have high traffic.

[128] For either Figure 13 or Figure 14, if the operator desires more information about any particular area or any particular area, a point-and-click interface allows the operator to navigate through representative still and video imagery of regions and/or activities that the system has detected and archived.

[129] Figure 15 illustrates another exemplary report for an aisle in a grocery store. The exemplary report includes an image from a video marked-up to include labels and trajectory indications and text describing the marked-up image. The system of the example is tasked with

searching for a number of areas: length, position, and time of a trajectory of an object; time and location an object was immobile; correlation of trajectories with areas, as specified by the operator; and classification of an object as not a person, one person, two people, and three or more people.

[130] The video image of Figure 15 is from a time period where the trajectories were recorded. Of the three objects, two objects are each classified as one person, and one object is classified as not a person. Each object is assigned a label, namely Person ID 1032, Person ID 1033, and Object ID 32001. For Person ID 1032, the system determined the person spent 52 seconds in the area and 18 seconds at the position designated by the circle. For Person ID 1033, the system determined the person spent 1 minute and 8 seconds in the area and 12 seconds at the position designated by the circle. The trajectories for Person ID 1032 and Person ID 1033 are included in the marked-up image. For Object ID 32001, the system did not further analyze the object and indicated the position of the object with an X.

[131] Referring back to block 22 in Figure 2, calibration can be (1) manual, (2) semi-automatic using imagery from a video sensor or a video recorder, or (3) automatic using imagery from a video sensor or a video recorder. If imagery is required, it is assumed that the source video to be analyzed by the computer system 11 is from a video sensor that obtained the source video used for calibration.

[132] For manual calibration, the operator provides to the computer system 11 the orientation and internal parameters for each of the video sensors 14 and the placement of each video sensor 14 with respect to the location. The computer system 11 can optionally maintain a map of the location, and the placement of the video sensors 14 can be indicated on the map. The map can be a two-dimensional or a three-dimensional representation of the environment. In

addition, the manual calibration provides the system with sufficient information to determine the approximate size and relative position of an object.

[133] Alternatively, for manual calibration, the operator can mark up a video image from the sensor with a graphic representing the appearance of a known-sized object, such as a person. If the operator can mark up an image in at least two different locations, the system can infer approximate camera calibration information.

[134] For semi-automatic and automatic calibration, no knowledge of the camera parameters or scene geometry is required. From semi-automatic and automatic calibration, a lookup table is generated to approximate the size of an object at various areas in the scene, or the internal and external camera calibration parameters of the camera are inferred.

[135] For semi-automatic calibration, the video surveillance system is calibrated using a video source combined with input from the operator. A single person is placed in the field of view of the video sensor to be semi-automatic calibrated. The computer system 11 receives source video regarding the single person and automatically infers the size of person based on this data. As the number of locations in the field of view of the video sensor that the person is viewed is increased, and as the period of time that the person is viewed in the field of view of the video sensor is increased, the accuracy of the semi-automatic calibration is increased.

[136] Figure 7 illustrates a flow diagram for semi-automatic calibration of the video surveillance system. Block 71 is the same as block 41, except that a typical object moves through the scene at various trajectories. The typical object can have various velocities and be stationary at various positions. For example, the typical object moves as close to the video sensor as possible and then moves as far away from the video sensor as possible. This motion by the typical object can be repeated as necessary.

[137] Blocks 72-25 are the same as blocks 51-54, respectively.

[138] In block 76, the typical object is monitored throughout the scene. It is assumed that the only (or at least the most) stable object being tracked is the calibration object in the scene (i.e., the typical object moving through the scene). The size of the stable object is collected for every point in the scene at which it is observed, and this information is used to generate calibration information.

[139] In block 77, the size of the typical object is identified for different areas throughout the scene. The size of the typical object is used to determine the approximate sizes of similar objects at various areas in the scene. With this information, a lookup table is generated matching typical apparent sizes of the typical object in various areas in the image, or internal and external camera calibration parameters are inferred. As a sample output, a display of stick-sized figures in various areas of the image indicate what the system determined as an appropriate height. Such a stick-sized figure is illustrated in Figure 11.

[140] For automatic calibration, a learning phase is conducted where the computer system 11 determines information regarding the location in the field of view of each video sensor. During automatic calibration, the computer system 11 receives source video of the location for a representative period of time (e.g., minutes, hours or days) that is sufficient to obtain a statistically significant sampling of objects typical to the scene and thus infer typical apparent sizes and locations.

[141] Figure 8 illustrates a flow diagram for automatic calibration of the video surveillance system. Blocks 81-86 are the same as blocks 71-76 in Figure 7.

[142] In block 87, trackable regions in the field of view of the video sensor are identified. A trackable region refers to a region in the field of view of a video sensor where an

object can be easily and/or accurately tracked. An untrackable region refers to a region in the field of view of a video sensor where an object is not easily and/or accurately tracked and/or is difficult to track. An untrackable region can be referred to as being an unstable or insalient region. An object may be difficult to track because the object is too small (e.g., smaller than a predetermined threshold), appear for too short of time (e.g., less than a predetermined threshold), or exhibit motion that is not salient (e.g., not purposeful). A trackable region can be identified using, for example, the techniques described in {13}.

[143] Figure 10 illustrates trackable regions determined for an aisle in a grocery store. The area at the far end of the aisle is determined to be insalient because too many confusers appear in this area. A confuser refers to something in a video that confuses a tracking scheme. Examples of a confuser include: leaves blowing; rain; a partially occluded object; and an object that appears for too short of time to be tracked accurately. In contrast, the area at the near end of the aisle is determined to be salient because good tracks are determined for this area.

[144] In block 88, the sizes of the objects are identified for different areas throughout the scene. The sizes of the objects are used to determine the approximate sizes of similar objects at various areas in the scene. A technique, such as using a histogram or a statistical median, is used to determine the typical apparent height and width of objects as a function of location in the scene. In one part of the image of the scene, typical objects can have a typical apparent height and width. With this information, a lookup table is generated matching typical apparent sizes of objects in various areas in the image, or the internal and external camera calibration parameters can be inferred.

[145] Figure 11 illustrates identifying typical sizes for typical objects in the aisle of the grocery store from Figure 10. Typical objects are assumed to be people and are identified by a

label accordingly. Typical sizes of people are determined through plots of the average height and average width for the people detected in the salient region. In the example, plot A is determined for the average height of an average person, and plot B is determined for the average width for one person, two people, and three people.

[146] For plot A, the x-axis depicts the height of the blob in pixels, and the y-axis depicts the number of instances of a particular height, as identified on the x-axis, that occur. The peak of the line for plot A corresponds to the most common height of blobs in the designated region in the scene and, for this example, the peak corresponds to the average height of a person standing in the designated region.

[147] Assuming people travel in loosely knit groups, a similar graph to plot A is generated for width as plot B. For plot B, the x-axis depicts the width of the blobs in pixels, and the y-axis depicts the number of instances of a particular width, as identified on the x-axis, that occur. The peaks of the line for plot B correspond to the average width of a number of blobs. Assuming most groups contain only one person, the largest peak corresponds to the most common width, which corresponds to the average width of a single person in the designated region. Similarly, the second largest peak corresponds to the average width of two people in the designated region, and the third largest peak corresponds to the average width of three people in the designated region.

[148] Figure 9 illustrates an additional flow diagram for the video surveillance system of the invention. In this additional embodiment, the system analyses archived video primitives with event discriminators to generate additional reports, for example, without needing to review the entire source video. Anytime after a video source has been processed according to the invention, video primitives for the source video are archived in block 43 of Figure 4. The video

content can be reanalyzed with the additional embodiment in a relatively short time because only the video primitives are reviewed and because the video source is not reprocessed. This provides a great efficiency improvement over current state-of-the-art systems because processing video imagery data is extremely computationally expensive, whereas analyzing the small-sized video primitives abstracted from the video is extremely computationally cheap. As an example, the following event discriminator can be generated: "The number of people stopping for more than 10 minutes in area A in the last two months." With the additional embodiment, the last two months of source video does not need to be reviewed. Instead, only the video primitives from the last two months need to be reviewed, which is a significantly more efficient process.

[149] Block 91 is the same as block 23 in Figure 2.

[150] In block 92, archived video primitives are accessed. The video primitives are archived in block 43 of Figure 4.

[151] Blocks 93 and 94 are the same as blocks 44 and 45 in Figure 4.

[152] As an exemplary application, the invention can be used to analyze retail market space by measuring the efficacy of a retail display. Large sums of money are injected into retail displays in an effort to be as eye-catching as possible to promote sales of both the items on display and subsidiary items. The video surveillance system of the invention can be configured to measure the effectiveness of these retail displays.

[153] For this exemplary application, the video surveillance system is set up by orienting the field of view of a video sensor towards the space around the desired retail display. During tasking, the operator selects an area representing the space around the desired retail display. As a discriminator, the operator defines that he or she wishes to monitor people-sized

objects that enter the area and either exhibit a measurable reduction in velocity or stop for an appreciable amount of time.

[154] After operating for some period of time, the video surveillance system can provide reports for market analysis. The reports can include: the number of people who slowed down around the retail display; the number of people who stopped at the retail display; the breakdown of people who were interested in the retail display as a function of time, such as how many were interested on weekends and how many were interested in evenings; and video snapshots of the people who showed interest in the retail display. The market research information obtained from the video surveillance system can be combined with sales information from the store and customer records from the store to improve the analysts understanding of the efficacy of the retail display.

[155] The embodiments and examples discussed herein are non-limiting examples.

[156] The invention is described in detail with respect to preferred embodiments, and it will now be apparent from the foregoing to those skilled in the art that changes and modifications may be made without departing from the invention in its broader aspects, and the invention, therefore, as defined in the claims is intended to cover all such changes and modifications as fall within the true spirit of the invention.